Required Data Centers and Interoperable services
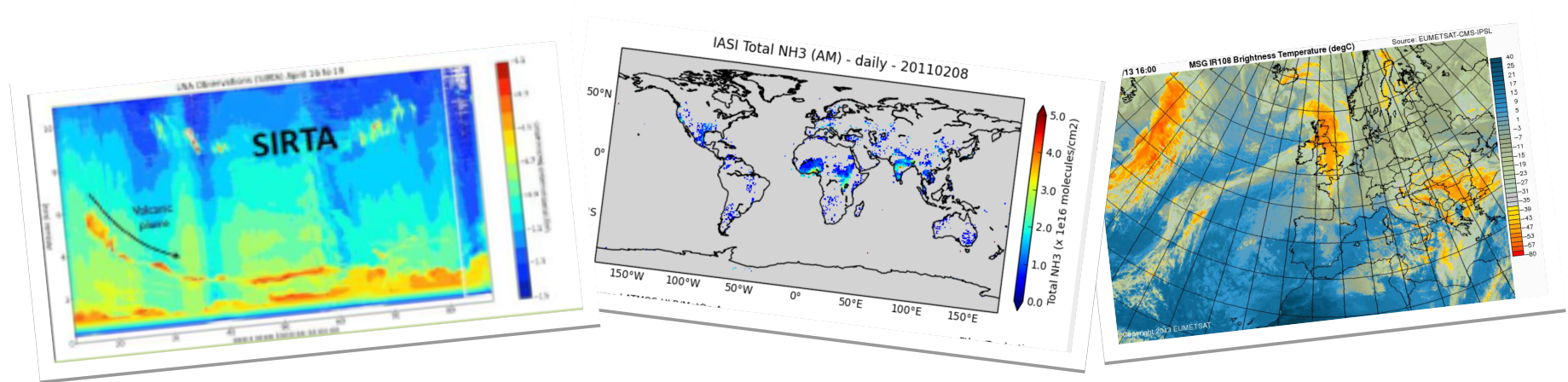
Sébastien Denvil

Institut Pierre Simon Laplace (IPSL)

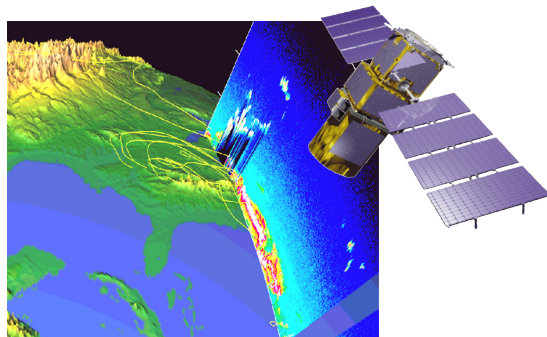Institut Pierre-Simon Laplace (**IPSL**) is a federation of 9 public research laboratories in the field of **Earth and planetary environmental sciences.**

www.ipsl.fr

Observ planets

IPSL wants to detect evolution and variation of climat and environment either from space or from ground stations that are part of international networks.

CALIPSO launched in april 2006 is part of the A-Train constellation. A platform dedicated to clouds and aerosols climatological studies.
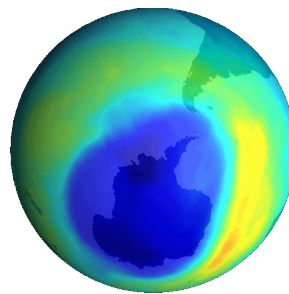
The ozone « hole »is continuously monitored from space since 1978

Ice are climate evolution witness. This ice core was extracted from the EPICA site on the Concordia basement in the Antartic.

Experiment to understand

IPSL scientists organize with their french and internationla partners multidisciplinary in situ observation campaign to understand key processes and interactions at play between natural environments (ocean, atmosphere, ice, continental surfaces)

ParisFog observational campaign aims to understand the physico-chemical processes controlling the fog life cycle.

The expansion of urban areas affects the quality of the air. Climate change and land use change are to fog-related settings. The campaign is led by the IPSL SIRTA remote sensing Site.

HyMeX campaign aim was to enhance hydrometeorologic risk predictability in the Mediterranean area.

ChArMEx campagn completed a comprehensive inventory of the atmospheric pollution in the Mediterranean area in order to improve our understanding atmospheric pollution and climat interactions.

Some Obs4MIPs basic tenets

- Use the CMIP5 simulation protocol as guideline for deciding which observations to select.
  Initial Target was monthly averaged (OMON, AMON) products on 1 x 1 degree grid

- Convert Observations to CMIP5 model output format
  CMOR output, NetCDF files, CF Convention Metadata, CMIP standard pressure levels, etc.
  Not a new product. Independent QC check before release.

- Includes a 6-8 page Technical Note describing strengths/weaknesses, uncertainties, caveats regarding comparisons with models.
  (at graduate student level)

- Available via ESGF
  (analogous to CMIP5)

Modelers

Observers

Analysis Community
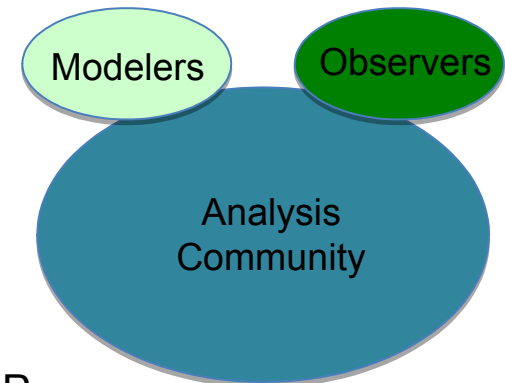
Obs4MIPs to report annually to WDAC/WCRP and WMAC/WCRP

Three independent initiatives

1) a US effort initiated and supported by NASA/JPL since about 2008 : Obs4MIPs_historical

2) a EU effort initiated by ESA since about 2010: Climate Change initiative (CMUG-CCI)

3) a joint EU & US effort initiated and supported by IPSL since about 2008: CFMIPObs (focus on Clouds)

Pursuing a similar objective:
Facilitate the use of satellite data for climate model evaluation

Using a similar approach:
Identify target quantities and make them available and easy to use by non remote sensing experts
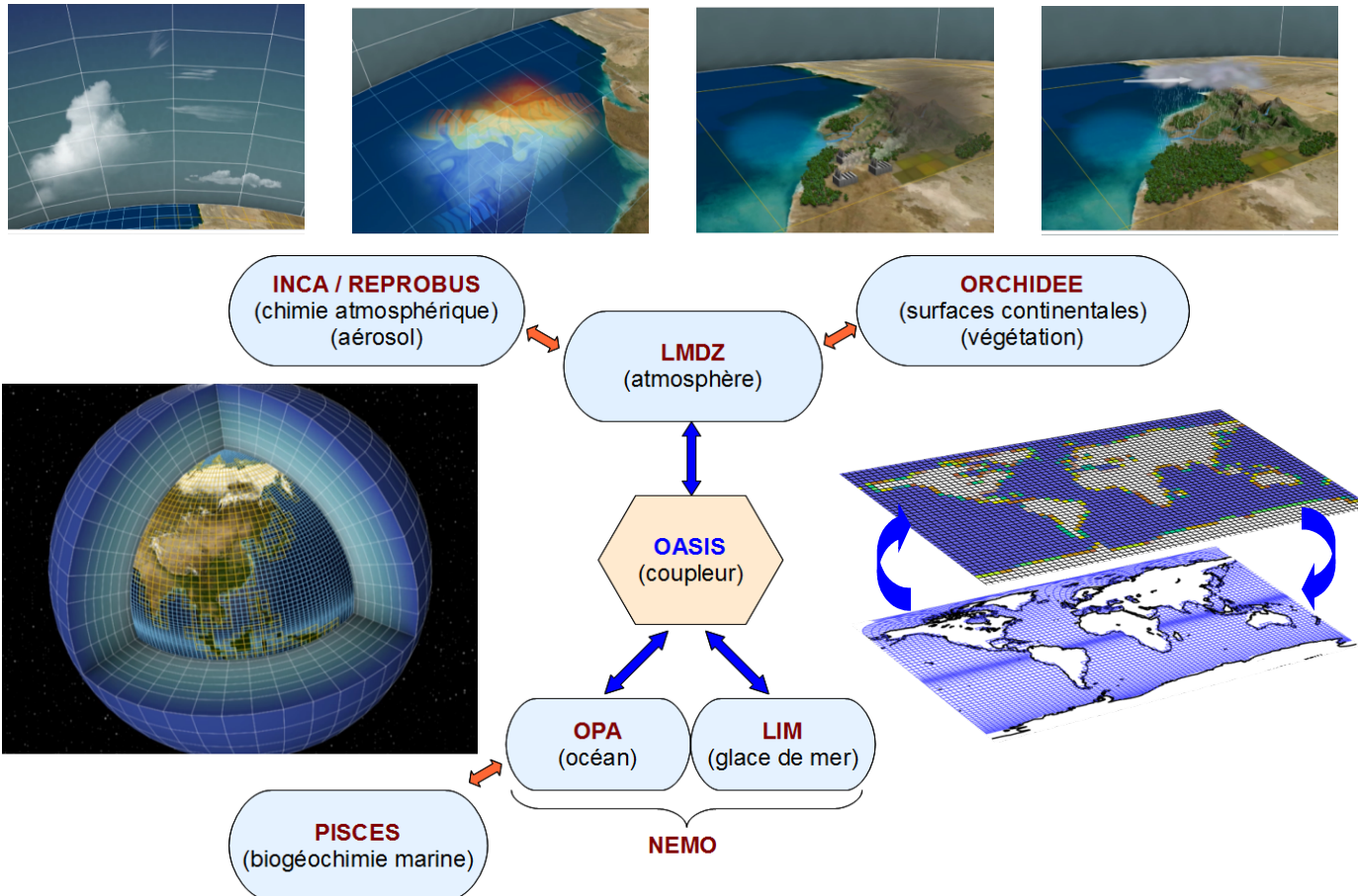
… To be merged together:
since 2012, Obs4MIPs_historical and CFMIP-Obs are available on the ESG under the name « Obs4MIPs »

=> Today EU contributes to Obs4MIPs with CFMIPObs through the IPSL node

Climate Models

The group's objective is the study of natural and anthropogenic variability in the global climate system. IPSL is also studying climate change impacts and usage of climate projections for adaptation to climate change related to industry. IPSL is one of the climate modeling centre of international repute contributing to the IPCC (Intergovernmental Panel on Climate Change).

INCA / REPROBUS
(chimie atmosphérique)
(aérosol)

ORCHIDEE
(surfaces continentales)
(végétation)

LMDZ
(atmosphère)

OASIS
(coupleur)

OPA
(océan)

LIM
(glace de mer)

NEMO
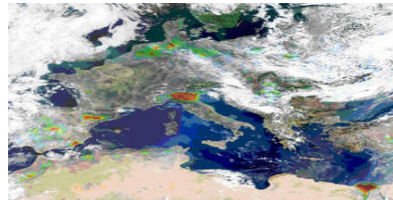
PISCES
(biogéochimie marine)

Crossing data

All data gathered together by IPSL, coming from field campaign, from observational network or from numerical simulations are stored in database. They are available to the scientific community within IPSL and at national and international level. Data are transfered to the civil society for operational applications (Climate Services, Copernicus program...).
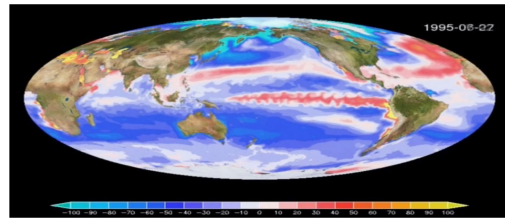
Models data
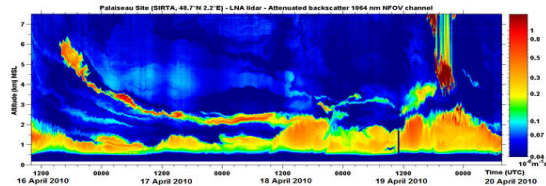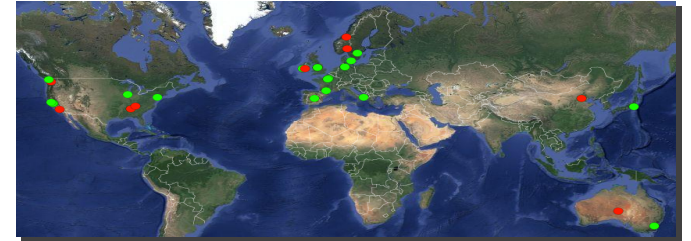
Ground observations

IPSL has engaged with ESGF developement and governance. IPSL is representing Météo-France, CERFACS and the National Supercomputing ecosystem within ESGF.

Satellite observations

IPSL mesoscale computing and data centre hosts data and computing services relevant for IPSL research.

Supercomputing Ecosystem

In France, created in 2007 GENCI aims is to catch up with european HPC standards.

- Represent France in PRACE
- End of 2014, national computing capacity was
  - 5,1 Pflop/s
  - 4 supercomputers with complimentary architecture
  - Two calls a year to grant access to the system

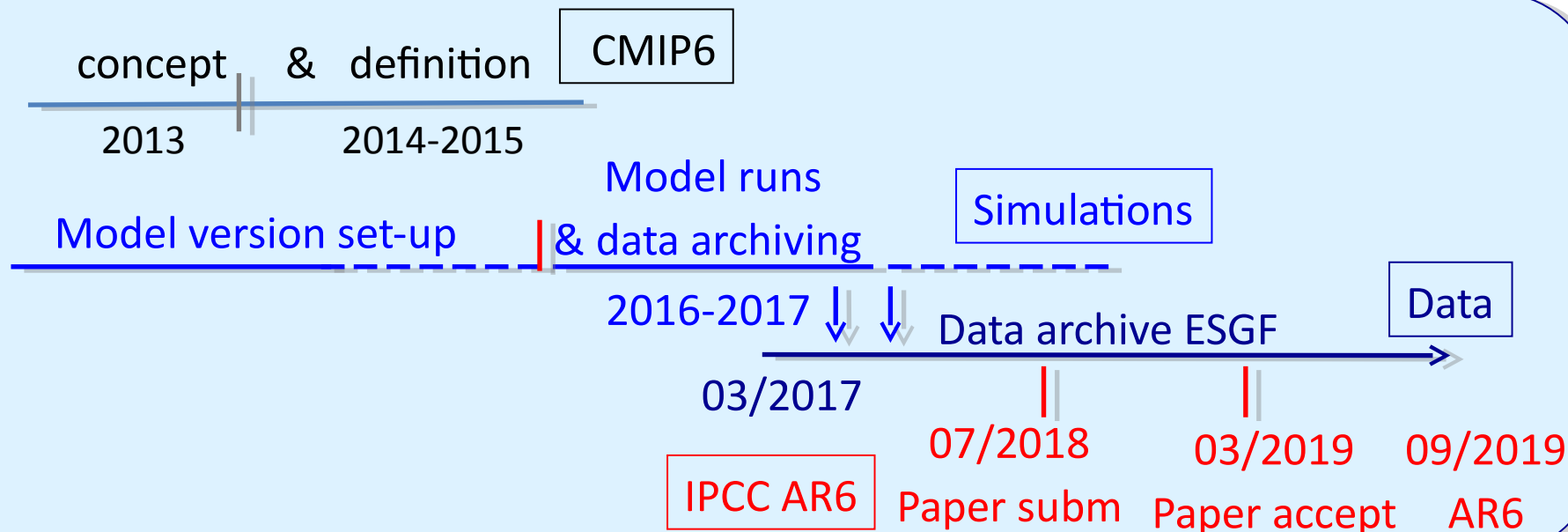| 2007	2014	2020
0 Pflop/s	≈20 Pflop/s	200 Pflop/s
0,02 Pflop/s	5,1 Pflop/s	40 Pflop/s
<0,01 Pflop/s	≈1,5 Pflop/s	5 Pflop/s
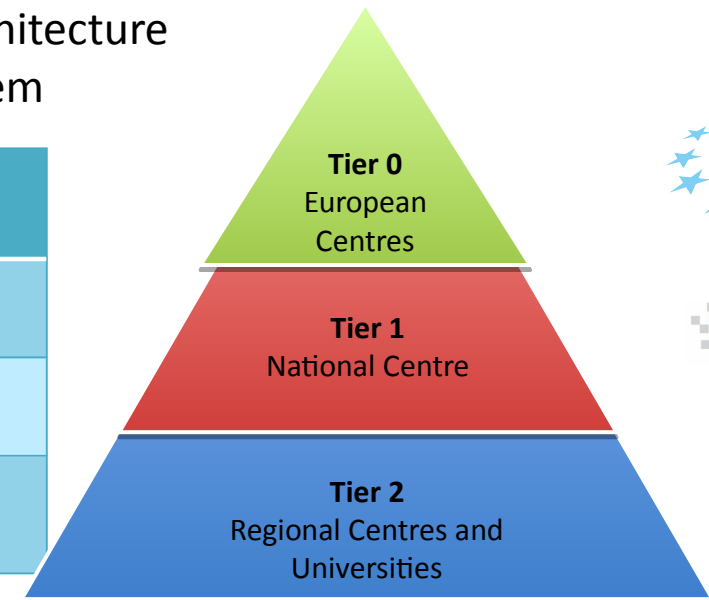
- CMIP6
  - 300 millions core hours 2016-2018
  - 14 Po (4 Po distributed on ESGF)
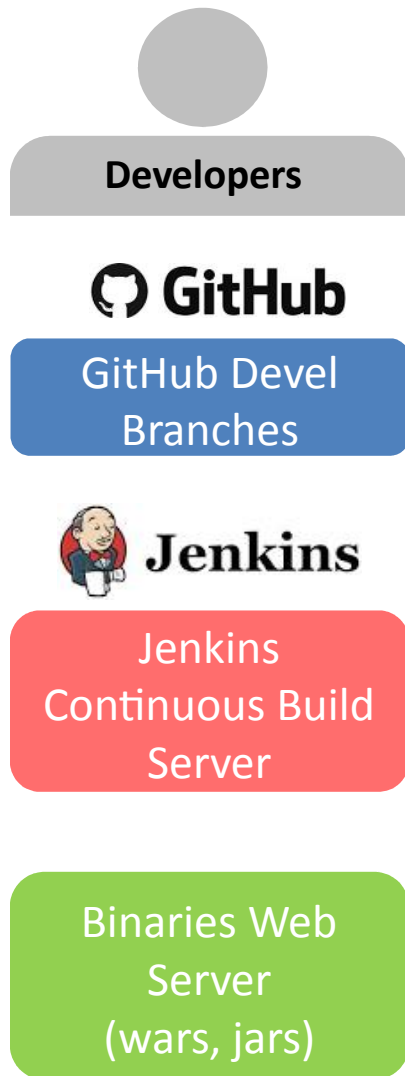 - CLIMERI : S. Joussaume, ESFRI

Roadmap

- **Data produced** by IPSL-CM model will be **distributed** from PRACE **Tier0** and GENCI **Tier1** supercomputing center (estimation : 4 Po).
- IPSL will host a CMIP6 **multi model replication** node on its **Tier2** center (a subset estimated to 4 Po).
- IPSL will provide a **national academic platform** to analyse CMIP6 outcomes on its **Tier2** center.
- IPSL (from its **Tier2** center) will provide access through iPython notebook to code and data that compose **particularly relevant figures**, especially figures from the IPCC report.

Roadmap

Coupled Model Intercomparison Project – CMIP6

CMIP6/AR6 cycle : International community under **strong** pressure

concept & definition | CMIP6 |

2013   2014-2015

Model runs

Model version set-up   & data archiving | Simulations |

2016-2017   Data archive ESGF | Data |

03/2017

07/2018   03/2019   09/2019

| IPCC AR6 | Paper subm   Paper accept   AR6

ESGF Continuous Build Platform

Developers

GitHub

Push Code

GitHub Devel Branches

Jenkins

Triggers Build

Jenkins Continuous Build Server

Publish Binaries if build completes

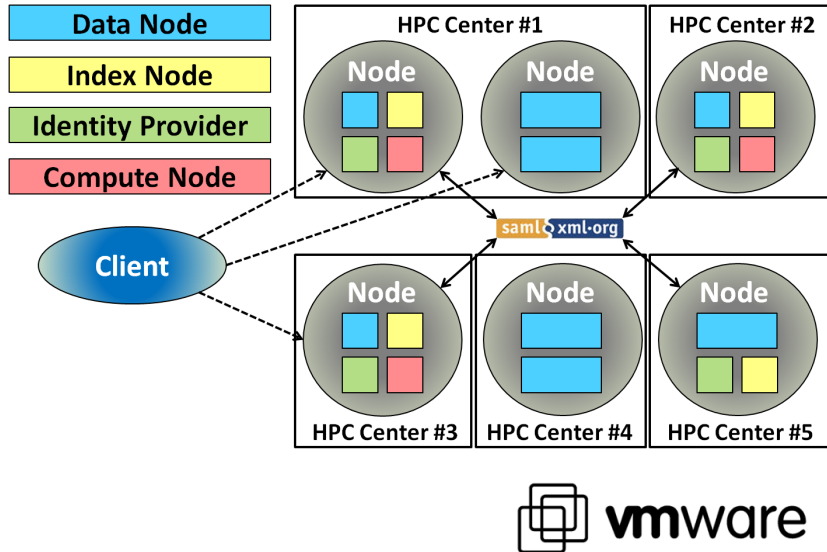Warning email if build breaks

Binaries Web Server (wars, jars)

@

Continuous Build
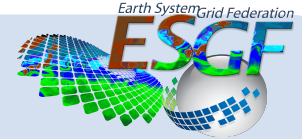
The ESGF software stack project source code is hosted on github repositories. 'Devel' branches are continuously updated with new features by development teams. The distribution binaries are then made available to the community for testing via a web server. Continuous build is useful to be aware of source code quality and inter project dependencies consistency in real time throughout development phases.

ESGF Test and Validation Platform

ESGF Test Infrastructure

- **Data Node**
- **Index Node**
- **Identity Provider**
- **Compute Node**

Client

HPC Center #1
Node **Node**

HPC Center #2
Node

saml xml·org

HPC Center #3
Node

HPC Center #4
Node

HPC Center #5
Node

vmware

ESGF Test Suite

- ➤ Python Nose - Test Framework
- ➤ Python Requests - HTTP Support
- ➤ Python Subprocess - System Execution
- ➤ Python Selenium - Browser Simulation
- ➤ Python Multiprocessing - Parallelisation

python

Integration Testing

The ESGF Test Federation is based on vmware virtual machines. It is completely independent from the production federation and is used to run the esgf test suite which performs integration tests in order to validate release candidates.

Earth System Grid Federation
ESGF
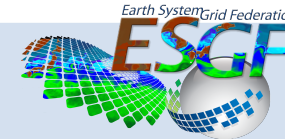
Replication and Versioning (DM + ESGF)

- Impact on CMIP6 data management DM) and ESGF governance (ESGF)
- Stable processes which are supervised by a board (the CDNOT Team) are needed for CMIP6 data consistency in ESGF
- CMIP6 data replication architecture:
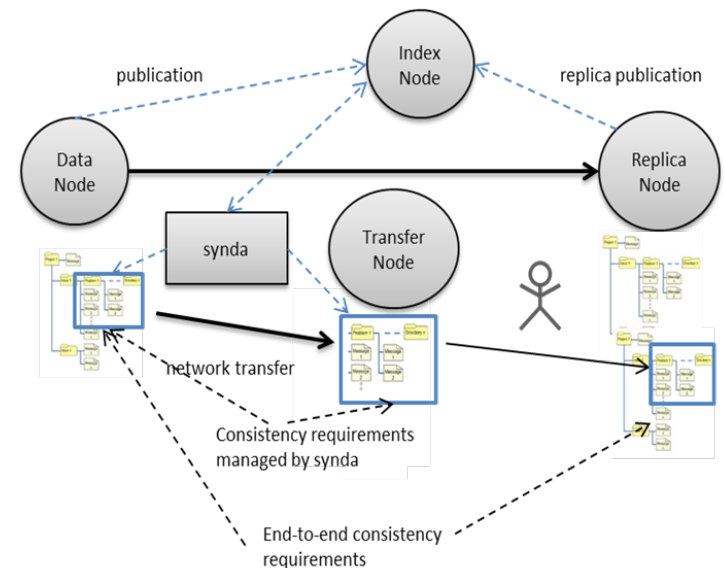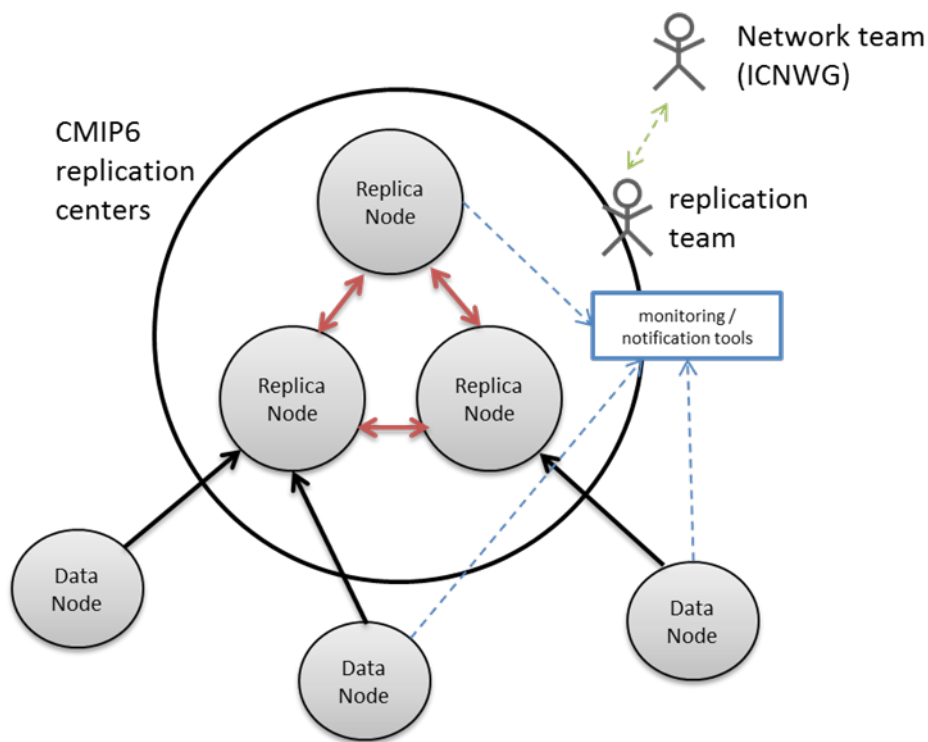
CMIP6 replication procedure

Data Analysis Environment

Mutualized : 200k€/yr shared funds.
Jointly *delivered* by
→IPSL laboratories.
Joint *users* (initially):
→IPSL community
Joint *users* (target):
→French Academic community

Analysis capabilities
Environmental Data
Compute Service
Web Service Provision for :
→Climate Science
→Earth Observation
→Environmental studies

Access services to ESGF System
Users don't have to find, download, and keep up to date the data they need

CMIP5,
CORDEX
Reanalysis,
Obs4MIPs
...

Big DATA Platform
Collaboration Environment
→ Access to Curated Archive.
→ Large shared "Group Work Spaces"
→ climate analysis enabled system
→ + 2.5 PB of high performance disk coupled to hundreds of cores configured for analysis

Summary

- We need not only to move computation to data, but aggregate our data collections

- Which required dedicated «HPD» platforms like Tier2 center.

- Security :
  - SELinux might become mandatory at some site !

- CMIP6 Challenges

  - CMIP6 satellite MIPs approach
    - « Decentralysed » science

    - « Decentralysed » data management

  - Replication, versioning, errata and network

  - Preparing for "near data processing" challenge
    - "download and process at home approach" for data analysis gets more and more problematic

Thank you for your attention